**Homework 6**

**Statistical Genomics**

**CROPS 545, Spring 2017**

Professor: Zhiwu Zhang

Due on April 28, 2017, Friday, 3:10PM, PST

**Data files**: mdp\_numeric.txt from GAPIT demo data. The data file can be download from <http://www.zzlab.net/GAPIT/GAPIT_Tutorial_Data.zip>. The data contains 281 individuals (row wise) and 3093 SNPs (column wise) coded as 0/1/2. The SNP ID, chromosome and position is indicated by a file named mdp\_SNP\_information.txt

**Hand in:** Email your report (PDF, limited to five page) and R source code (text file) with subject of “CROPS545 HW6” to Zhiwu.Zhang@WSU.edu. Name your files as following:

Homework6\_ firstname\_lastname.pdf and Homework6\_ firstname\_lastname.R

**Grade components**: 1) Hypothesis or statement; 2) Results; 3) Methods; 4 presentation; 5) R source code (clarity, simplicity and documenting comments)

**Objectives**: 1) validation; 2) invalidate validation; 3) cross validation; 4) gBLUP; and 5) ridge regression.

1. Use GAPIT.Phenotype.Simulation function to simulate phenotypes with heritability of 50% controlled by 20 QTNs having effects with standard normal distribution. Display the distribution of QTN effects, and the correlation between the total genetic effects (breeding values) and phenotypes of individuals (5 points).
2. Perform GWAS on the simulated phenotypes with all the individuals by using FarmCPU and selected the top 20 associated SNPs. Perform random division of all the individuals into two even (roughly) sub populations A and B. Estimate the effects of the 20 associated markers in sub population A. Use the estimated effects of the 20 SNPs to predict the phenotypes and BV in sub population B. Repeat the random division 30 times. Report the means and standard deviations of the prediction accuracy (20 points).
3. Repeat (2) except randomly shuffling the simulated phenotypes before GWAS. Describe the difference from (2) and your expectation (15 points).
4. With the simulated phenotypes from (1), randomly select 80% of the individuals as training population and the rest as testing population. Perform gBLUP with GAPIT. Calculate the correlations between the predictions and phenotypes, and the correlation between predictions and breeding values in training and testing populations separately. Repeat the random selection and prediction 30 times. Compare the means and standard deviations of the correlations in training and testing population (20 points).
5. With the simulated phenotypes from (1), perform random division of the population into five even (roughly) sub populations (5-folds). Perform 5-folds cross validation to evaluate accuracy in testing population by using Ridge regression with rrBLUP. Calculate accuracy as the average of the correlations between predicted and observed phenotypes, and the correlations between predicted and observed breeding values in the testing populations. Repeat the random division and prediction 30 times. Compare the mean and standard deviation of the accuracy with the result from (4) (20 points).